Running head: NO FREE LUNCH

Is There a Free Lunch In Inference?

Jeffrey N. Rouder

University of Missouri

Richard D. Morey

University of Groningen

Josine Verhagen

University of Amsterdam

Jordan M. Province

University of Missouri

Eric-Jan Wagenmakers

University of Amsterdam

Jeff Rouder

rouderj@missouri.edu

Abstract

The field of psychology, including cognitive science, is vexed in a crisis of confidence. Although the causes and solutions are varied, we focus here on a common logical problem in inference. The default mode of inference is significance testing which has a *free lunch property* where researchers need not make detailed assumptions about the alternative to test the null hypothesis. We present the argument that there is no free lunch, that is, valid testing requires that researchers test the null against a well-specified alternative. We show how this requirement follows from the basic tenets of conventional and Bayesian probability. Moreover, we show in both the conventional and Bayesian framework that not specifying the alternative may lead to rejections of the null hypothesis with scant evidence. We review both frequentist and Bayesian approaches to specifying alternatives, and show how such specifications improve inference. The field of cognitive science will benefit as consideration of reasonable alternatives will undoubtedly sharpen the intellectual underpinnings of research.

Keywords: inference, philosophy of science, statistics, replication crisis

Is There a Free Lunch In Inference?

Prelude: The Tragedy of Sally Clark

In fields such as medicine and law, statistical inference is often a matter of life or death. When the stakes are this high, it is crucial to recognize and avoid elementary errors of statistical reasoning. Consider, for instance, the British court case of Sally Clark (Dawid, 2005; Hill, 2005; Nobles & Schiff, 2005). Clark had experienced a double tragedy: her two babies had both died, presumably from cot death or sudden infant death syndrome (SIDS). If the deaths are independent, and the probability of any one child dying from SIDS is roughly 1/8543, the probability for such a double tragedy to occur is as low as $1/8543 \times 1/8543 \approx 1$ in 73 million. Clark was accused of killing her two children, and the prosecution provided the following statistical argument as evidence: Because the probability of two babies dying from SIDS is as low as 1 in 73 million, we should entertain the alternative that the deaths at hand were due not to natural causes but rather to murder. And indeed, in November 1999, a jury convicted Clark of murdering both babies, and she was sentenced to prison.

Let us reflect on what happened. Forget the fact that the deaths may not be independent (due to common environmental or genetic factors), suppress any worries about how the rate of 1/8543 was obtained, and focus solely on the inferential logic used in the case. Assuming that the probability of two babies dying from SIDS is indeed as low as 1 in 73 million, to what extent does this incriminate Clark? The prosecution followed a line of statistical reasoning similar to that used throughout the empirical sciences: one postulates a single hypothesis (i.e., the null hypothesis: "Sally Clark is innocent") and then assesses the unlikeliness of the data under that hypothesis. In this particular case, the prosecution felt that the probability of two babies dying from SIDS was sufficiently

small to reject the null hypothesis of innocence, and thereby accept the hypothesis that Sally Clark is guilty.

The flaw in the reasoning above is the consideration of only one hypothesis: the null hypothesis that Sally Clark is innocent and her babies died from SIDS. Hence, the only datum deemed relevant is the low background probability of two babies dying from SIDS. But what of the background probability of two babies dying from murder? In 2002, President of the Royal Statistical Society Peter Green wrote an open letter to the Lord Chancellor in which he explained that "The jury needs to weigh up two competing explanations for the babies' deaths: SIDS or murder. The fact that two deaths by SIDS is quite unlikely is, taken alone, of little value. Two deaths by murder may well be even more unlikely. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation." (Nobles & Schiff, 2005, , p. 19). Statistician Phil Dawid (2005, p. 8) agreed: "(...) if background evidence of double-death-rates due to SIDS (or other natural causes) is relevant, then surely so too should be background evidence of double-death-rates due to murder. To present either of these figures without some assessment of the other would be both irrational and prejudicial." Ray Hill (2005, p. 15) showed how different the statistical conclusions are when one takes into account both sides of the coin: "Obtaining reliable estimates based on limited data is fraught with difficulty, but my calculations gave the following rough estimates. Single cot deaths outnumber single murders by about 17 to 1, double cot deaths outnumber double murders by about 9 to 1 and triple cot deaths outnumber triple murders by about 2 to 1. (...) when multiple sudden infant deaths have occurred in a family, there is no initial reason to suppose that they are more likely to be homicide than natural."

After being tried, convicted, and serving three-years in prison, Sally Clark was acquitted and freed. She subsequently developed serious psychiatric problems and died three years later from acute alcohol poisoning.[1]

The statistical error that plagued the Sally Clark case also besets statistical inference in the empirical sciences. To see why, consider a revision of the letter by Peter Green in which we have replaced the terms that are specific to the Sally Clark case with more general statistical concepts: "The researcher needs to weigh up two competing explanations for the data: The null hypothesis or the alternative hypothesis. The fact that the observed data are quite unlikely under the null hypothesis is, taken alone, of little value. The observed data may well be even more unlikely under the alternative hypothesis. What matters is the relative likelihood of the data under each hypothesis, not just how unlikely they are under one hypothesis."

Statistical Inference in Psychology and the Desire for a Free Lunch

In psychology, statistical inference is generally not a matter of life or death. Nonetheless, large-scale adoption of methods that focus on the null alone without recourse to well-specified alternatives does have deleterious long term consequences that are becoming ever more apparent. Inferential logic that is ill-suited for determining the truth of Sally Clark's guilt will be equally ill-suited to finding the answer to any other question.

Recent work suggests that psychological science is facing a "crisis of confidence" (Yong, 2012; Pashler & Wagenmakers, 2012) fueled in part by the suspicion that many empirical phenomena may not replicate robustly (Carpenter, 2012; Roediger, 2012; Yong, 2012). Note that psychology shares this crisis of confidence with other fields; for instance, pharmaceutical companies recently complained that they often fail to replicate preclinical findings from published academic studies (Begley & Ellis, 2012; Osherovich, 2011; Prinz, Schlange, & Asadullah, 2011), with some replication rates as low as 11%. Another concern is that reputable journals have recently published findings that are highly implausible, the most prominent example featuring a series of studies on extra-sensory perception (Bem, 2011; Storm, Tressoldi, & Di Risio, 2010).

The current crisis of confidence has many causes (Nosek & Bar–Anan, 2012; Nosek, Spies, & Motyl, 2012), but one contributing cause is that the field is relying on flawed and unprincipled methods of stating evidence, and our everyday practices need revision. Just as in medicine and law, one of the main goals of psychological scientists is to assess and communicate the evidence from data for competing positions. For example, in the simplest hypothesis-testing framework, the analyst is trying to assess the evidence for the null and alternative hypotheses. The prevailing approach used in the field is what we term the "$p < .05$ rule", that is, effects are considered to be demonstrated if the associated p values are less than .05. Most researchers hold complicated views about the wisdom of this rule. On one hand, most of us realize that data analysis is more thoughtful and organic than blind adherence to the $p < .05$ rule. On the other hand, many would argue that the $p < .05$ rule has served the community well as a rough-and-ready safeguard from spurious effects (Abelson, 1997). Researchers by and large seem to believe that although the $p < .05$ rule is not perfect in all applications, it is useful as a conventional guide for experimenters, reviewers, editors, and readers.

We argue here that the $p < .05$ rule does not serve us well, and in fact, its use contributes to the crisis. At its core, the $p < .05$ rule assumes what we term here the *free lunch of inference*. The free lunch refers to the ability to assess evidence against the null hypothesis without consideration of a well-specified alternative. In significance testing, for instance, the analyst does not need to commit to any particular alternative to assess the evidence against the null. A contrasting approach is inference with power analyses. With a power analysis, a researcher can choose an appropriate balance between Type I and Type II errors. But to do so, the alternative must be specified in detail: the effect size under the alternative is set to a specified value.

Most of us would rather have a free lunch. We may have the intuition that our conclusions are more valid, more objective, and more persuasive if it is not tied to a

particular alternative. This intuition, however, is wrong. We argue here that statistical evidence may be properly understood and quantified only with reference to detailed alternatives. Moreover, and more importantly, inference with and without an alternative yield markedly different results. Inference without judicious consideration of alternatives is capriciously tilted toward concluding that there are effects in data, and those who adopt it require too low a standard of evidence to make bona-fide claims. Specifying alternatives certainly requires more thought and effort, but it will lead to more principled and more useful evaluations of results. As the saying goes, "there ain't no such thing as a free lunch".[2]

The message that there is no free lunch in testing originates from a seminal *Psychological Review* article by Edwards, Lindman, & Savage (1963). This paper correctly identified the methodological problems we are dealing with today, and suggested the appropriate remedy: namely, that lunch must be paid for. Edwards et al. was largely ignored by psychologists, but has been influential in statistics through the work of Berger, Raftery, and others (e.g., Berger & Delampady, 1987; Raftery, 1995, 1999; Sellke, Bayarri, & Berger, 2001). The arguments from Edwards et al. have recently been reimported back to psychology by Dienes (2008), Gallistel (2009), Rouder, Speckman, Sun, Morey, & Iverson (2009), and Wagenmakers (2007), among others. There is sage wisdom in Edwards et al. that remains relevant, important, and under-appreciated; the reader will benefit from a careful rereading of this classic. One might consider our message to be a half-century commemoration of Edwards et al.'s ideas and impacts. A less statistical argument is provided by Platt (1964) in his classic piece introducing *strong inference*. The first element of strong inference according to Platt is "devising alternative hypotheses" (p. 347).

Our argument that well-specified alternatives are needed in testing is motivated by logic and probability. We start with the strong logic of falsification, in which alternatives

are not needed. We note that extending strong falsification to the probabilistic case fails, and then show how one must specify detailed alternatives under both frequentist and Bayesian probability frameworks. To make the case using frequentist probability, we rely on the fact that sure knowledge is defined in the large-data limit; for example, as data collection continues, sample means converge to true means. Frequentist testing should be perfect in the large-sample limit; that is, with an infinite amount of data, testing should provide *always* for the correct decision. This perfection in the large-sample limit is called *consistency* and, unfortunately, as will be discussed, significance testing is not consistent. Consistent frequentist testing is possible, but it requires that analysts specify an alternative. Bayesian probability implies a different constraint on inference: prior beliefs must be updated rationally in light of data. This constraint leads immediately to inference by Bayes factor, and the Bayes factor requires priors. It is through these priors that the analyst specifies alternatives. The message is that regardless of whether one is a frequentist or Bayesian, in principle one must specify alternatives.

The Need for Alternatives: Starting from Logic

Logical inference provides the foundation for scientific inference. Propositional logic offers a means for rejecting strong theories through *modus tollens*. For example, if a theory predicts a certain experimental result could not occur, and the result does occur, then the theory may be rejected; i.e.,

(Premise) If Hypothesis H is true, then Event X will not occur.

(Premise) Event X has occurred.

(Conclusion) Hypothesis H is not true.

Here, note that we do not need to posit an alternative to Hypothesis H. The above argument is valid, and the conclusion is justified if the premises are certain. Yet, in almost

all cases of empirical inquiry, we are not certain in our premises, and when there is noise or error contraindicative events may occur even when Hypothesis H holds. When premises involve uncertainty, we may invoke probability theory. The modified version of the argument above adapted for significance testing is

(Premise) If Hypothesis H is true, then Event X is unlikely.

(Premise) Event X has occurred.

(Conclusion) Hypothesis H is probably not true.

In this argument, the impossibility of Event X in the first premise has been replaced by rarity: that is, the event will *usually* not occur, assuming the Hypothesis $H$is true. In the case of NHST, the rare event is observing that the p value is lower than a certain level, usually .05. This argument, however, is NOT valid. The conclusions do not follow, and belief in the conclusion is not justified by belief in the premises. Pollard and Richardson (1987) demonstrate the invalidity of the argument by the following example:

(Premise) If Jane is an American, then it will be unlikely that she is a U. S. Congressperson.

(Premise) Jane is a U. S. Congressperson.

(Conclusion) Jane is probably not an American.

The argument is obviously invalid, and yet it is exactly the same form as the NHST argument used through the sciences (Cohen, 1994; Pollard & Richardson, 1987). The reason why it is invalid is that it fails to consider the alternative to the hypothesis "Jane is an American." If Jane were not an American, the observed event – that she is a Congressperson – would be impossible. Far from calling into doubt the hypothesis that Jane is an American, the fact that she is a Congressperson makes it certain that she is an American. Because the form of the above argument is invalid, the argument underlying

NHST is also invalid, and for exactly the same reason. Alternatives matter for valid inference; there is no free lunch.

In the following sections, we further develop the argument that there is no free lunch from the basic definitions of probability. This development shows not only the need for alternatives but how alternatives may be specified and integrated into inference. When alternatives are considered, inference changes appreciably—often more evidence is needed to claim that there is an effect. The proper interpretation of this fact is that inference with a free lunch overstates effects, and, by extension, that cognitive scientists – and all other users of NHST – have sometimes been claiming effects with scant evidence.

The Frequentist Argument For Alternatives

Consistency

Following the lead of Fisher (1925, 1955), significance tests are most often used without consideration of alternatives. The logic of significance testing described in the previous section is often called "Fisher's disjunction": either the null hypothesis is false, or a rare event has occurred. Since rare events typically do not occur, one can supposedly infer that the null hypothesis is probably false. Fisher's philosophical rival, Jerzy Neyman, opposed the testing of null hypotheses without reference to an alternative. Neyman (1956) goes so far as to state with respect to frequentist hypothesis tests that "the main point of the modern theory of testing hypotheses is that, for a problem to make sense, its datum must include not only a hypothesis to be tested, but in addition, the specification of a set of alternative hypotheses..."

In the frequentist paradigm, probabilities are defined as long-run proportions. The probability of an event – say, that a coin lands heads up – is the proportion of times a coin lands heads up in the limit of infinitely many flips. This so-called large sample limit can be used to assess the adequacy of frequentist methods. In small data sets, analysis will be

error prone due to random variation. But as sample size increases, good frequentist methods become more accurate, and in the large-sample limit, they reach the correct answer to arbitrary precision. This property is called *consistency*.

Because the central tenet of frequentist probability is convergence in the large sample limit, it is important that hypothesis testing and model comparison be consistent, that is, they should reach always the correct answer in the large-sample limit. From a frequentist point-of-view, consistency is a minimal property for good inference. Is significance testing consistent? If it is we expect to make the correct decision in the large-sample limit regardless of whether the null is true or false. First consider the case that the null hypothesis is false. In this case and in the large-sample limit, the null hypothesis will always be correctly rejected, which is consistent behavior. The difficulty arises when the null is true. In this case, by construction, test statistics will lead to a rejection with a set Type I error rate, usually 5%. The problem is that this probability does not diminish with sample size; regardless of how large one's sample is, the analyst is condemned to make Type I errors at the preset level. These errors, however, violate consistency. The situation is difficult, significance testing, which is based on the frequentist definition of probability, violates the core frequentist notion that knowledge is certain in the large-sample limit

It may seem that these violations are small and inconsequential—after all, what are the consequences of a 5% error rate? To appreciate some of these consequences, consider the study by Galak, LeBoeuf, Nelson, & Simmons (2012); in seven experiments, the authors tried to replicate the ESP results reported in Bem (2011). The seventh and last experiment featured $2,469$ participants, and the result yielded $t(2468) = -0.23$, $p = .59$. These results are clearly not significant, and it may be tempting to conclude that the statistical results from such an extremely large sample must be highly dependable. But this is a mistake – given that the null hypothesis is true and ESP does not exist, there was

a 5% chance of obtaining a significant result. In other words, regardless of how many participants one tests, whether ten or ten million, with a fixed α level of .05 there is a 5% chance of falsely rejecting a true null hypothesis. So collecting more data does not increase one's chances of drawing the correct conclusion when the null hypothesis is true. In fact, using significance testing for high-powered replication experiments where the null hypothesis is true resembles a game of Russian roulette. In the case of Experiment 7 from Galak et al. (2012), the field of psychology dodged a bullet that had a 5% chance of hitting home ("Psychologists prove existence of ESP: results significant with 2,469 participants!").

Gaining Consistency Through Specification of Alternatives

To ameliorate the problem, we could make a seemingly slight change in significance testing to assure consistency. Instead of letting α be fixed at .05, we let it vary with sample size such that, in the limit, $\alpha \to 0$ as $N \to \infty$. We expand the notation, and let α_N denote the Type I error rate at a particular sample size. In significance testing, bounds are set such that $\alpha_N = .05$ for all N. Consider just a modest variant of the standard approach where we set the critical bound such that $\alpha_N = \min(.05, \beta_N)$, where β_N is the Type II error rate at sample size N (the Type II error rate is the probability of failing to detect the alternative when it is true, and is the complement of power). Here we never let the Type I error exceed .05, just as in the standard approach, but we also decrease it as the sample size increases so that the Type I error rate never exceeds the Type II error rate. With this schedule, both α_N and β_N necessarily decrease to zero in the large sample limit, and consequently inference is consistent. This new rule meets two requirements: (1) It is clearly consistent, as both Type I and Type II error rates decrease to zero in the large sample limit; and (2) Type I errors should not be any more plentiful than 1-in-20 if the null is indeed true.

This variant approach differs from $p < .05$ significance testing in a few crucial

aspects. One is that it requires the analyst to compute β_N. The computation of β_N, the Type II error rate, in turn requires a specification of an alternative effect size (just as power does). For the purpose of demonstration, we set ours to .4 in value. Here, we see that to get consistent inference wit this approach, the analyst must specify an alternative. This specification is indeed paying for lunch.

Although this variant rule is similar to the original $p < .05$ rule, the resulting conclusions may differ markedly. Displayed in Figure 1 are the critical effect-size values needed to reject the null, and these are plotted as a function of sample size. The wide, light-colored line is for standard significance testing where $\alpha_N = .05$. As can be seen, the needed critical effect size falls with sample size, and will fall to zero in the large sample limit. The dashed line shows the critical effect size for $\alpha_N = \min(.05, \beta_N)$ rule for an alternative effect size of .4 in value. Critical effect sizes reach a bound of .2 and decrease no further. With this small variation, we have eliminated researchers' ability to reject the null with small effects; indeed effects need to be larger than .2 to reject the null. The value of .2 is no coincidence, and it is easy to show that this value is half the value of the specified alternative (Dienes, 2014; Faul, Erdfelder, Lang, & Buchner, 2007). We also computed critical effect sizes for the rule that Type II errors are 5 times as large as Type I errors ($\alpha_N = \beta_N/5$; see the thin solid line). This rule leads to consistent inference as both Type I and Type II errors decrease to zero in the large sample limit. It too requires that the value of effect size be set under the alternative. Adopting this rule leads to the same result that small effects cannot be used as evidence to reject the null no matter the sample size. Th value of the alternative matters. It determines a floor value on effect size for significance. .This specification, which has practical impact, is made before data are collected and should reflect experience, context, and the theoretical questions at hand. Principled frequentist inference here is surprisingly subjective.

The idea of balancing Type I and Type II errors to achieve consistency is not new.

Neyman and Pearson (1933) suggested that researchers strike some sort of balance between the two sorts of errors, based on the research context. Recently several authors have proposed on various approaches to adjusting α as a function of sample size so that α approaches 0 in the large-sample limit (see Perez & Pericchi, 2014). All the proposed approaches that we are aware of require the analyst to provide additional information, that is, they have to pay for lunch.

.

Confidence Intervals

Although hypothesis testing is nearly ubiquitous, there have been repeated calls for replacing hypothesis testing with estimation. The proponents of estimation suggest that researchers report point estimates of parameters of interest (e.g., effect size) and confidence intervals (CIs) (Cumming, 2014; Grant, 1962; Loftus, 1996). Most researchers use CIs to exclude implausible values, and they may reject the null if the CI does not cover zero (Hoekstra, Morey, Rouder, & Wagenmakers, 2014).

Consider the behavior of a researcher who computes the sample mean and draws the associated 95% confidence interval to determine which values are plausible. Whatever the true value is, and however large the sample size, the confidence interval will have the same 5% probability of excluding the true value, even in the large sample limit (Neyman, 1937). That is, confidence intervals are inconsistent! Needed is a plan of increasing the coverage index, say from 95% to 100%, as the sample size increases. To our knowledge, researchers do not increase the coverage index in large samples.

The Bayesian Argument for Alternatives

Bayesian Basics

In the Bayesian framework, probability describes a degree of belief. It is a subjective concept from first principles, and different analysts are expected to arrive at different probabilities for events depending on their background knowledge. The key principle in Bayesian statistics is that beliefs, represented as probability, should be revised optimally in light of data. The steps to optimal revision are well-known—the analyst simply follows Bayes' rule. Perhaps the primacy of Bayes' rule in Bayesian analysis is stated most directly by Bradley Efron in his presidential address to the American Statistical Association (Efron, 2005). Efron writes, "Now Bayes' rule is a very attractive way of reasoning, and fun to use, but using Bayes' rule doesn't make one a Bayesian. *Always* using Bayes' rule does, and that's where the practical difficulties begin." (p. 2, italics in original).

The focus of Bayesian analysis is not what to believe but rather how to change beliefs in light of data. Consider a priming example where each of I participants responds in a primed and control condition. Let Y_i be the difference score across the two conditions for the ith person. The following model is placed on the difference scores:

$$Y_i \sim \text{Normal}(\mu, \sigma^2). \tag{1}$$

It is helpful to work in a common unit, effect size, $\theta = \mu/\sigma$. For simplicity in exposition, let's assume σ^2 is known.[3] With this simplification, the model becomes

$$Z_i \sim \text{Normal}(\theta, 1), \tag{2}$$

where $Z_i = Y_i/\sigma$ are standardized difference scores.

The goal is to update beliefs rationally about the effect size θ in light of data. Before observing the data, the analyst's beliefs are expressed as a probability distribution. For instance, the analyst may believe that effect sizes tend to cluster around zero and

rarely exceed 2.0 in magnitude. These beliefs are expressed as a distribution called the *prior distribution*, and denoted $\pi(\theta)$. Figure 2A, green dashed line, shows a prior distribution on effect size. The graph is that of a *probability density function*, and the probability that θ is in any interval is the area under the curve for that interval.

Now the data are observed, and these are shown as red tick marks near the x-axis. The analyst uses Bayes' rule, presented subsequently, to rationally update her beliefs. This updated distribution is called the *posterior distribution*, and it expresses beliefs conditional on observing specific data. The probability density function for the posterior is the solid blue line in Figure 2A. This density is denoted $\pi(\theta|\boldsymbol{Z})$ where $\boldsymbol{Z}$ are the data.

Bayes' rule, which provides for an optimal updating of beliefs (see Jeffreys, 1961), is

$$\pi(\theta \mid \boldsymbol{Z}) = U(\theta, \boldsymbol{Z}) \times \pi(\theta).$$

The term $U(\theta, \boldsymbol{Z})$ is the updating factor that tells the analyst how to update the prior, $\pi(\theta)$, to reach the posterior, $\pi(\theta|\boldsymbol{Z})$. The updating factor is:

$$U(\theta, \boldsymbol{Z}) = \frac{f(\boldsymbol{Z} \mid \theta)}{f(\boldsymbol{Z})}.$$

The numerator of the updating factor, $f(\boldsymbol{Z}|\theta)$ is the probability density function of the data for a specific parameter value. This probability density is often straightforward to assess and is proportional to the likelihood function for the model. It is the *conditional probability density of the data*, conditional on a parameter value. The denominator, $f(\boldsymbol{Z})$, is the marginal rather than conditional probability of the data. It is the probability density of data averaged across *all* parameter values. It is found by integration, and the relationship is known as the Law of Total Probability (Jackman, 2009):

$$f(\boldsymbol{Z}) = \int f(\boldsymbol{Z} \mid \theta)\pi(\theta)\,\mathrm{d}\theta.$$

Hence, the updating factor is the probability of data at the evaluated parameter value relative to the probability density of data averaged across all parameter values. The

updating factor emphasizes parameter values that are more likely than the average and attenuates parameter values that are less likely.

Figure 2B, on the lower left, shows how updating works. The probability of the data conditional on θ is the black line and that averaged across all parameter values is the flat red line. Let's consider a single point: $\theta = 1$ (circles). The conditional value at that point is 1.2×10^{-6}, which is 2.9 times greater than the marginal (average) value of 4.0×10^{-7}. The updating factor at this point is therefore 2.9, and the posterior probability is 2.9 greater than the prior for $\theta = 1$. Also shown is the value at $\theta = 0$ (squares). Here the data are less probable for this value than average by a factor of 3.3, and indeed this point is 3.3 times less plausible after seeing the data than before.

Bayes' rule may be written as a single equation. A useful form is

$$\frac{\pi(\theta \mid \boldsymbol{Z})}{\pi(\theta)} = \frac{f(\boldsymbol{Z} \mid \boldsymbol{\theta})}{f(\boldsymbol{Z})}. \tag{3}$$

The left-hand side of the equation describes how the data lead to a revision of belief about a particular parameter value, and we may call this the *evidence* from the data. The right-hand side describes how probable the data are for a particular parameter value relative to how probable they are on average. The phrase "probability of data" is best thought of as *prediction*. When the observed data are relatively probable, then the model has predicted them well; when they are relatively rare, the model has predicted them poorly. The right-hand side is the relative in predicting the data for a given parameter value relative to an average across all parameters. If the observed data are predicted by the parameter value more so than average, then evidence for the parameter value is increased. Likewise, if the observed data are predicted less well by a parameter value than average, the evidence for the parameter value is decreased. **The deep meaning of Bayes' rule is that evidence is prediction.**

The preceding usage of Bayes' rule showed how beliefs about parameters in models

should be updated. Yet, we are often interested in comparing models themselves. In the current case, we are interested in a null model where $\theta = 0$, and whether this model or the previous model better describes the data. Let $\mathcal{M}_1$ be our previous effects model, which may be stated as

$$\begin{aligned} Z_i \mid \theta &\sim \text{Normal}(\theta, 1), \\ \theta &\sim \text{Normal}(0, 1). \end{aligned}$$

The first line is sometimes called the *data-generating step* as it relates data to parameters. The second line is the prior. The combination of both statements is the model.

The data are multidimensional, but fortunately for this problem the sample mean is a sufficient statistic. Therefore, we can restrict our attention to $\bar{Z}$ and be assured the same results as considering all the data. Figure 2C, solid line, shows $f(\bar{Z})$, the marginal probability of the statistic $\bar{Z}$. We may view this graph as a prediction of the model for $\bar{Z}$. The distribution here is proper and the density integrates to 1.0.

Suppose we wish to compare this general model to a null model, denoted $\mathcal{M}_0$. The null model has the same data-generating mechanism but a prior where all the mass is on a single point of $\theta = 0$. It may be stated as:

$$\begin{aligned} Z_i \mid \theta &\sim \text{Normal}(\theta, 1). \\ \theta &= 0. \end{aligned}$$

Here, the difference between the models is the difference between the priors. This focus on priors as defining different points of theoretical interest is an advantage of Bayesian model specification as discussed by Vanpaemel (2010).

The predictions for $\bar{Z}$ from the null model are given in Figure 2C as a dashed line. With these sets of predictions from the null and general models, we are ready to perform model comparison. Let's consider the data in Figure 2A, which are shown in the small

tick-marks on the x-axis. The value of the mean is $\bar{Z} = .725$. The vertical lines in Figure 2C and 2D are for this sample-mean value. The predictive density for this value are .091 under the null and .300 under the general model (see Figure 2C). The ratio is 3.3-to-1 in favor of the general model over the null. Had the observed sample size been small, say .10, the null model would have predicted it better, that is, the ratio would have favored it relative to the general model.

In Figure 2D, we plot the ratios of the predictions with the null model in the numerator and the general model in the denominator. Because the null model predicts smaller sample-mean values better than the general model, the ratio is > 1 for these smaller values. Conversely, because the general model predicts larger sample-mean values better than the null model, the ratio is < 1 for these larger values. From Figure 2C, we computed the ratio at 3.3-to-1 in favor of the general. The reciprocal of 3.3-to-1 is .305, which is the value at the sample mean in Figure 2D. These ratios of predictive densities are called *Bayes factors*.

An application of Bayes' rule shows the important role of the Bayes factor. In Bayesian analysis, beliefs may be placed on models themselves. Prior beliefs, those held before considering data, for Models $\mathcal{M}_1$ and $\mathcal{M}_0$ are denoted $\Pr(\mathcal{M}_1)$ and $\Pr(\mathcal{M}_0)$, respectively. Bayes' rule is used to update these beliefs:

$$\Pr(\mathcal{M}_1|\boldsymbol{Z}) = \frac{f(\boldsymbol{Z}|\mathcal{M}_1)\Pr(\mathcal{M}_1)}{f(\boldsymbol{Z})}, \quad \Pr(\mathcal{M}_0|\boldsymbol{Z}) = \frac{f(\boldsymbol{Z}|\mathcal{M}_0)\Pr(\mathcal{M}_0)}{f(\boldsymbol{Z})},$$

where $\Pr(\mathcal{M}_1|\boldsymbol{Z})$ and $\Pr(\mathcal{M}_0|\boldsymbol{Z})$ are posterior beliefs. It is convenient to express these posterior beliefs on models as odds:

$$\frac{\Pr(\mathcal{M}_1|\boldsymbol{Z})}{\Pr(\mathcal{M}_0|\boldsymbol{Z})} = \frac{f(\boldsymbol{Z} \mid \mathcal{M}_1)}{f(\boldsymbol{Z} \mid \mathcal{M}_0)} \times \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_0)}. \tag{4}$$

The term $f(\boldsymbol{Z} \mid \mathcal{M}_1)/f(\boldsymbol{Z} \mid \mathcal{M}_0)$ is the *Bayes factor*, and it describes the updating factor or the extent to which the data cause revision in belief (Kass & Raftery, 1995). Once

again, evidence, the degree to which belief change about two models, is the relative predictive accuracy of the models.

We denote the Bayes factor by B_{10}, where the subscripts indicate which two models are being compared. A Bayes factor of $B_{10} = 10$ means that prior odds should be updated by a factor of 10 in favor of the effects model $\mathcal{M}_1$; likewise, a Bayes factor of $B_{10} = .1$ means that prior odds should be updated by a factor of 10 in favor of the null model $\mathcal{M}_0$. Bayes factors of $B_{10} = \infty$ and $B_{10} = 0$ correspond to infinite support from the data for one model over the other with the former indicating infinite support for model $\mathcal{M}_1$ and the later indicating infinite support for model $\mathcal{M}_0$. Bayes' rule has the same interpretation as before—the updating factor on beliefs, the evidence in data, is the relative predictive accuracy of the models (Wagenmakers, Grünwald, & Steyvers, 2006). Because Bayes factors index the evidence for models from data, their use has been recommended repeatedly for psychological research (e.g., Edwards et al., 1963; Gallistel, 2009; Rouder et al., 2009; Wagenmakers, 2007).

Specifying Alternatives

One of the key properties of the Bayesian approach is that it is dependent on the specification of the alternative. In the previous example, we specified the alternative through the prior, $\theta \sim \text{Normal}(0, 1)$. It is a reasonable prior that does not assume the direction of the effect and seems broadly appropriate for psychological research where effect sizes tend to be small. It is worthwhile to explore how this specification affects Bayesian analyses.

Figure 3A shows three effects model priors: the original one plus one twice as variable and one half as variable. The null is also presented as a spike at zero. One way of thinking of these models is in terms of flexibility. The narrowest prior is the least flexible model because it is compatible only with small effects whereas the widest prior is

compatible with a wider range of effects. The predictions of these three priors as well as the predictions of the null model is shown in Figure 3B. The predictive distribution for the null model is most near zero, and the probability densities spread as the priors become more flexible. The Bayes factors are the ratio of predictions between the null model and the effects models and serve as the evidence for the null model relative to an effect model. Three Bayes factor curves, one for each prior, are shown in Figure 3C.

Consider the differences for a sample mean of $\bar{Z} = 0$. The data are best predicted by the null and least well predicted by the effects model with the widest prior. The Bayes factor depends on the prior of course, and the value is about twice as big for the null relative to the narrowest prior compared to the null relative to the widest prior. We view this dependency as desirable. The widest prior is most flexible, it is a less parsimonious account of small sample sizes than a narrower prior or a null model. If a researcher's alternative specifies big effects yet small effects are observed, these small effects are more compatible with the null than with the specified alternative.

A credible interval approach

Although many Bayesian statisticians advocate Bayes factors as a principled measure of evidence derived directly from Bayes' rule, this advocacy is far from universal. In psychology, methods that emphasize interval estimates rather than Bayes factor ratios have become popular (Kruschke, 2011, 2012). The posterior distribution of the parameter may be characterized by its *credible interval.* The credible interval is conceptually similar to a confidence interval, and covers the middle 95% of the posterior. Figure 4 shows an example of a wide prior (dashed line), and a well-localized posterior (solid line) with a relatively narrow 95% posterior credible interval (lightly shaded area between vertical lines).[4] Kruschke recommends basing inference about the characteristics of this interval. There are a few versions of Kruschke's recommendations but all have the following

characteristics: If the credible interval is sufficiently narrow and localized well away from zero, then the null point has low plausibility. Likewise, if the credible interval is sufficiently narrow to be in a small interval around zero then the null point has high plausibility. Kruschke recommends focusing on the narrowness of the credible interval, and gathering enough data so that the credible interval is below a preset criterial width, say .25 on the effect-size scale. In Figure 4, the posterior is well localized at an effect size of .165 and the credible interval is somewhat narrow with a width of .22. This smallish credible interval excludes zero, and , consequently, the interpretation may be that the null has low plausibility.

Kruschke's approach is heuristic inasmuch as it does not follow directly from Bayes' rule. We could, however, use Bayes' rule directly to compute the amount evidence for the point zero. In fact, the consideration of the density at zero before and after seeing the data is a perfectly valid way to compute the Bayes factor between the null model and the effects model used in estimation (This fact was first published by Dickey,1971, though he attributes it so Savage, and the computation method is known as Savage-Dickey ratios.) Before seeing the data, the null is not especially plausible, and the density value at zero is .1261 (see the red point in Figure 4). After the data, however, the density has decreased by the slightest factor to .1256, that is, the data had virtually no effect aon the plausibility of the zero value. Indeed, the Bayes factors between the models is just about 1.0. After seeing the data, we should have the same belief in the null, the zero point, as we did before seeing the data.

Here we have a divergence. By using posterior credible intervals we might reject the null, but by using Bayes' rule directly we see that this rejection is made prematurely as there is no decrease in the plausibility of the zero point. Updating with Bayes' rule directly is the correct approach because it describes appropriate conditioning of belief about the null point on all the information in the data.

Summary

The Bayes factor has many advantages: it allows researchers to state evidence for or against models, even nulls, as dictated by the data, it provides a report of evidence without recourse to reject or fail-to-reject, avoiding the dichotomous thinking inherent in significance testing (Cumming, 2014), and provides a natural penalty for model complexity (Myung & Pitt, 1997). Nonetheless, we think the most important advantage is that it is principled—it is the only approach that meets the Bayesian obligation to update rationally. In this case, beliefs about models themselves are updated rationally through Bayes' rule. And, in turn, Bayes' rule may be seen as equating evidence and prediction. Other methods of stating evidence for or against positions necessarily violate Bayes' rule and are not ideal. If one updates through Bayes' rule, then the specification of the alternative through the prior is important and consequential.

How to Specify the Alternative: A Real-World Example

We have argued that researchers must commit to specific alternatives to perform principled inference. Many researchers, however, may feel unprepared to make this commitment. We understand the concern. The good news is that making judicious commitments is not as difficult as it might seem, because researchers have more information than they may realize (Dienes, 2011, 2014). The argument goes as follows:

The key to specification of the alternative is consideration of effect-size measures, which are widely understood and can be used in almost all cases. Importantly, effect sizes have a natural scale or calibration. For example, an effect size of 10.0 is very large, and effects of this size would be obvious with just a handful of observations. Likewise, effect sizes of .01 are too small to be discriminated in most experiments, in fact, trying to do so would exhaust our subject pools. The four panels on the left side of Figure 5 captures these constraints. In all four panels, the priors have common properties: 1. the prior is

symmetric about zero reflecting ignorance about the direction of the effect; 2. there is decreasing plausibility with increasing effect size. The difference between them is in the scale or range of effect sizes, and these may be seen in the x-axis values. The prior labeled "Too Wide," shows a specification of the alternative that makes a commitment to unrealistically large values of effect size. Evidence for the null would be overstated in this case because the alternative is too broad. There are three additional panels, labeled"Wide", "Narrow," and "Too Narrow," that show other specifications. The specification "Too Narrow" shows commitments to very small effect sizes, and would not be useful in most applications in experimental psychology. The priors "Wide" and "Narrow" define the outer points of reasonable specification—priors more dispersed than "Wide" seem too dispersed, and priors more concentrated than "Narrow" seem too concentrated. These specifications are referenced by a single quantity, the scale of effect size, denoted σ_δ, which ranges in these specifications from 10, for "Too Wide," to .02, for "Too Narrow." A reasonable range for σ_δ is from .2, the narrow prior, to 1, the wide prior. In practice, we often use $\sigma_\delta = .5$ and $\sigma_\delta = \sqrt{2}/2$, depending on our prior expectations.

These different specifications will lead to different Bayes factors, but the degree of variation is far less than one might expect. The right side of Figure 5 shows the effect of the prior scale of the alternative, σ_δ, on Bayes factor for three values of the t-statistic for a sample size $N = 50$. The highlighted points in green show values for $\sigma_\delta = .2$ and $\sigma_\delta = 1$, which define a reasonable range. The red points are from scales considered too extreme. Prior scale does matter, and may change the Bayes factor by a factor of 2 or so, but it does not change the order of magnitude. The priors used on the left side of Figure 5 with specified scale on effect size serves as an example of a *default prior*, a prior that may be used broadly, perhaps with some tuning across different contexts. We have recommended default priors in our advocacy of Bayes factors (Morey & Rouder, 2011; Morey et al., 2013; Rouder et al., 2009; Rouder et al., 2012; Rouder et al., 2013; Rouder & Morey 2012;

Wetzels, Grassman, & Wagenmakers, 2012; Wetzels & Wagenmakers, 2012), and we implement them for t-tests, regression, and ANOVA, as well in our web applets (`pcl.missouri.edu/bayesfactor`) and BayesFactor package for `R` (Morey & Rouder, 2014).

The specification of normals surrounding the null is not necessary. Dienes (2011) and Gallistel (2009) recommend alternative subjective priors that may be centered where the researcher expects the effect size to be. For example, a research in a Stroop experiment might suspect a relatively large effect size and center the prior there. In this regard, these priors capture more contextual information. Another class of priors that are becoming popular in Bayesian analysis are what is termed "nonlocal priors" (Johnson & Rossell, 2012). These priors are used as an alternative to the null, and they specify alternatives that are located quite far form the null for maximal contrasts.

For concreteness, we illustrate the Bayes factor approach with an example from the moral reasoning literature. One of the critical paradigms in the literature is the *runaway trolley problem*: A runaway trolley car with brakes that have failed is headed towards five workers on the train track. Participants are put into the shoes of a bystander with the option to pull a lever which redirects the train onto a side track. Unfortunately, there is a worker on this side track as well. The bystander has two choices: either do nothing and let five workers be killed or pull the lever and actively kill one worker. A critical question in this literature is what the results reveal about moral reasoning. Hauser, Cushman, Young, Jin, & Mikhail (2007) argue that the results reveal underlying universal moral competencies (the ability to sacrifice one to save five) while Rai & Holyoak (2010) question whether the paradigm could disentangle universal competencies from heuristic-type biases famously documented by Tversky & Kahneman (1974).

To test whether framing biases could mask moral competencies, Rai & Holyoak (2010) asked participants to declare reasons for pulling the lever and killing the one and

saving the five. We consider two possible outcomes. The first is that listing reasons could prime a more utilitarian judgment of pulling the lever. In one condition, participants were asked to list only two reasons, in the other condition they were asked to list as many reasons as they could, up to seven reasons. The prediction is that those in the seven-reason condition would be more utilitarian in their moral judgments than those in the two-reason condition, that is, they would more readily actively sacrifice one worker to save five workers.

Rai & Holyoak (2010) considered a second, alternative theory from consumer psychology. In consumer psychology, giving more reasons for a choice often decreases rather than increases the probability of making that choice (Schwarz, 1998). The reason for this apparent paradox is straightforward—as people attempt to list more reasons, they either fail to do so or list reasons of lower quality than their primary reasons. A focus on either this failure or on these reasons of lower quality results in a less favorable view of the choice. This explanation, called here the *framing alternative*, predicts that asking participants for many reasons for pulling the lever results in a lower probability of a moral judgment to do so.

We consider the results of Rai & Holyoak (2010), Experiment 1, where 124 participants first provided reasons for pulling the lever, and then indicated on a 4-point scale whether they agreed with the judgment of pulling the lever. Half of the participants were given the two-reason instructions, the remaining half were given the seven-reason instructions. The critical contrast is whether participants agreed with pulling the lever more in the two-reason than seven-reason condition. Before assessing this critical contrast, Rai and Holyoak first assessed whether the instruction manipulation was effective in generating more reasons. Participants provided an average of 3.2 and 1.7 reasons, respectively, in the seven- and two-reason conditions, respectively. The corresponding t-value, 5.46, is large and corresponds to a small p-value well under .01. To assess the

evidence in a principled fashion, we compared the null to a wide default alternative with $\sigma_\delta = 1$ (see Figure 5), and the resulting Bayes factor is about 50,000-to-1 in favor of the alternative. Therefore, there is compelling evidence that the instruction manipulation did indeed affect the numbers of reasons provided.

The critical contrast is whether participants agreed with pulling the lever more in the two-reason than seven-reason condition. Rai & Holyoak (2010) report more agreement with pulling the lever in the two-choice condition than in the seven-choice condition (2.17 vs. 1.83 on the 4-point agreement scale). The t value for the contrast is $t(122) = 2.11$, which is just significant at the .05 level. Based on this significance, the authors conclude, "We found that people are paradoxically less likely to choose an action that sacrifices one life to save others if they are asked to provide more reasons for doing so."

To perform a Bayesian analysis, we chose the following three models: The first is the null—there is no effect of instructions on the moral judgements—which is certainly plausible. We chose two different alternatives: the one-sided alternative that judgments are greater in the seven-choice condition, and the other one-sided alternative from the framing alternative that judgments are less in the seven-choice condition. The prior scale on effect size in these alternatives is $\sigma_\delta = .7$, an in-between value reflecting that effect sizes in framing manipulations and moral-reasoning dilemmas tend to vary. Relative to the null, the Bayes factors are 15.1-to-1 against the utilitarian priming hypothesis and 2.77-to-1 in favor of the framing alternative (again, there is more agreement observed in two-reason than seven-reason condition). The evidence from the data stand in opposition to the utilitarian priming explanation as 15-to-1 is a sizable revision in beliefs, but they do not provide nearly as much support for the framing alternative. A less than a 3-to-1 revision from a full-sized experiment is not very convincing to us.

Conclusion

In this paper we have argued that if one is to test hypotheses, or more generally compare models, then as a matter of principle one must pay the price of specifying a reasonable alternative. Moreover, we argue that paying the price for principled inference may have material effects on conclusions. When alternatives are reasonable, small observed effects in large samples do not readily serve as evidence for true effects. In this section, we comment on why specifying alternatives, that is paying for lunch, would lead to a better psychological science.

Is Specifying Alternatives Too Subjective?

The $p < .05$ rule purportedly provides field-wide protection: supposedly, when we, the scientific community, insist on $p < .05$ rule, we implement a safeguard against which individual researchers cannot exaggerate their claims. Our view is that the field will give up no safeguards whatsoever by adopting a specification viewpoint. In fact and to the contrary, having researchers state *a priori* their expectations of effect sizes under the alternative will vastly improve the transparency and integrity of analysis. Analyses will be more transparent because researchers will need to state and defend their choices of the alternative, and researchers and readers can easily evaluate these defenses much as they evaluate all other (subjective) aspects of research including the operationalization of variables and the link between experimental results and theoretical objectives. In this regard, analysis is put on an equal footing with all other aspects of research. Of course, analysts must invest some thought into possible alternatives and defend their choices. We believe that psychological scientists are up to the task, and when they do so, they will draw better inferences from data about theory.

Hypothesis Testing vs. Estimation

Some critics have objected to hypothesis testing on the grounds that the null model is never true (Cohen, 1994; Meehl, 1978). Those who advocate this position recommend that researchers replace hypothesis testing with estimation of effect sizes (Cumming, 2014). There are three separate issues associated with effect-size advocacy: that the null is never true, that estimation should replace testing, and that whether estimation requires no commitment to nulls or alternatives. We discuss them in turn:

We find the position that "the point null is never true" difficult on several grounds (Iverson, Wagenmakers, & Lee, 2010; Morey & Rouder, 2011; Rouder & Morey, 2012). However, it is not necessary to consider point nulls to show the necessity of specifying alternatives. Each of our examples may be restated using interval null hypotheses, such as $-.1 < \mu < .1$, without any loss. The same inferential logic that applies to point nulls also applies to bounded intervals. The "null is never true" argument does not obviate the need to specify alternatives.

A second issue is whether estimation should replace testing as is advocated by Cumming (2014). We have no issue with estimation as part of the scientific process. However, there are questions that estimation cannot answer; for instance, "Does the Higgs Boson exist?" "Do all humans come from a single individual?" Or, pedestrianly, "How concordant is a certain set of data with a certain theory?" These questions are about evidence for hypotheses (Morey, Rouder, Verhagen, & Wagenmakers, 2014). Although estimation has been suggested as a replacement for testing for several decades now, testing remains nearly ubiquitous. Perhaps testing retains its popularity because it answers questions that researchers ask. Estimation fails because researchers do not wish to simply catalog effect sizes across tasks and manipulations. They want to understand them in theoretical terms (Jeffreys, 1961, Morey et al., 2014; for a charming metaphoric account of the same point, see Forscher, 1963).

.

Finally, there is the question whether estimation is truly model free. We use sample values to "estimate" true values. True values, however, are not natural things, but mathematical parameters in a model. And the choice of model affects what is the best estimator. If two researchers do not agree upon the model, then they won't agree on the best estimate. In our case, the consideration of the possibility of lawfulness and regularity often changes what might be the best estimate. Consider the following example inspired by Good (1983):

A wealthy donor has decided to test the skills of a data analyst. The donor tells the analyst that she will donate $10,000 to the analyst's favorite charity if the analyst can predict exactly the number of heads in 1000 flips of a coin. If the analyst cannot predict the exact number, the donor will deduct an amount from the total according to a schedule: she will deduct $1 from the total if the analyst is off by a single count; $4 if the analyst is off by two counts, $9 if the analyst is off by three counts and so on. The donor warns the analyst that the coin may or may not be fair. To compensate for this complication, the donor allows the analyst to observe 1000 flips to form an opinion, and then has to predict the outcome of the next 1000 flips. We consider three different possible outcomes of the first 1000 flips. **a.** Suppose there are 750 heads in the first 1000. We should be certain that the coin is not fair and the best prediction is that another 750 heads will occur in the next 1000 flips. Here it seems as if specification of models is unnecessary. **b.** Suppose instead that there were 510 heads in the first 1000 flips. It seems likely that this value is far more concordant with a fair coin than with an unfair coin, and if this is the case, the better prediction might be 500. The 10-head deviation from 500 in the first set is likely noise and if this is the case, then predictions based on it will be more error prone than those at the 500 value. Indeed, if the unfair-coin-alternative is the hypothesis that all probabilities are equally likely, the outcome of 510 heads supports the

fair coin null over this alternative by a factor of 20-to-1. **c.** Perhaps the most interesting case is if there are 540 heads in the first 1000. This value is intermediate, and in fact, is equally probable under the null as under the above alternative. The best estimate now is the average of 500 and 540, the estimates under each equally-plausible model, which is 520. The analyst had to specify the alternative to obtain it.

The above example with the donor and coin is not far fetched. Let's take an example such as whether short-term exposure to a violent video game causes increased subsequent aggression compared to exposure to a nonviolent video game (Anderson & Dill, 2000; Ferguson et al., 2008). On one hand, it is *a priori* plausible that violent-video game exposure increases aggression. On the other, it is *a priori* plausible that participants are well aware that the violence in the game is not real and, consequently, it has no psychological impact whatsoever. This null is an invariance—-aggression is invariant to the degree of violence in video-games. Let's suppose a researcher observes a sample effect size (Cohen's d) of .06., and we wish to estimate the true effect size. If we believe the alternative is the correct model after seeing the data, then we should stick to the .06 value. But if we believe that the null is the correct model after seeing the data, then 0 is the appropriate value. The best estimate depends on our degree of belief in each possible model after seeing the data, and it will be between 0 and .06.

Figure 6 shows the best estimate for a sample size of 50 observations and an alternative model where the prior on effect size is a standard normal. Let's take the case of large sample effect sizes, say those greater than .5 or less than -.5. The posterior probability of the null is small and the best effect-size estimate is close to the sample effect size. Conversely, for small effect sizes, say those between -.1 and .1, the posterior probability of the null is large and the true effect-size estimate is greatly attenuated. This attenuation of magnitude is known as shrinkage and is considered a benefit of Bayesian modeling (Efron & Morris, 1977). For the case of an observed effect size of .06, the

weights on the null is 87%, and when zero is mixed in at this weight, the estimated effect size reduces .0078. This value is less than 1/13th of the original value. It is our view that many of the small effect sizes reported in the literature are overstatements of the true value because researchers have neglected to consider the possibility that the null may hold. Typical estimates are predicated on an effect occurring, and that predication may be premature in most settings.

The critical point is that principled estimation is not model-free, and researchers who make commitments are acting with more principle than those who do not. Researchers may choose to use summary statistics like sample means. They should note, however, that this choice assumes an effects model where the probability that the true effect is null is itself zero. Researchers who make such a choice should be prepared to defend this logical consequence, though, in truth, it seems difficult to defend in experimental situations.

Bayesian Probability Works Well In Scientific Contexts

Bayesian probability has many theoretical and practical advantages. In the Bayesian view, probabilities are the expression of belief rather than long-run proportions. Probabilities are held by observers rather than physical properties of the system, and can be changed or updated rationally in light of new information (de Finetti, 1974). The notion that probabilities are mutable beliefs corresponds well with scientific reasoning where opinions are revised through experiments (Rozenboom, 1960). Because Bayesian analysis has the built-in concept of prior belief, it is natural and transparent to specify alternatives. Moreover, unlike most testing approaches, Bayesian analysis does not require decision or action. One does not commit to the dichotomy of rejecting a hypothesis or failing to do so. Instead, Bayesian analysis is *epistemic* in that the analyst can state updating factors—how their beliefs changed—without making any decisions or asking the

readers to do so as well. The ability to report fine grained continuous measures of evidence seems compatible with a more enlightened science Cumming (2014).

The Adverse Effects of Free-Lunch Myth

In the beginning of this paper, we argued that the free-lunch myth—the belief that we may learn about the null without specifying alternatives—has led to a mind-set and culture that is academically counterproductive. From a pragmatic perspective, free-lunch inference overstates the evidence against the null and leads to rejections with too low a threshold. This low threshold fools researchers into thinking their results are more reliable than they truly are, much as the jury rejected the null hypothesis about Sally Clark's innocence. Yet, we are equally concerned with a broad perspective. Significance testing has reduced data analysis to a series of prescribed procedures and fixed decision rules. Inference has become an intellectual dead zone that may be done by algorithms alone (Gigerenzer, 1998). We need only provide *pro forma* evaluation of the outputs. And in doing this *pro forma* evaluation, we abandon our responsibility to query the data. We use significance testing to reify results without appreciation of subtlety or qualification, and to manufacture a consensus even when this consensus is unwarranted or meaningless. We live down to the $p < .05$ rule, and, in this process divorce what we may know about our data from what we can tell others about them. It is in this space, where hypothesis testing is simultaneously a bureaucratic barrier and the sanctioned ritual of truth, that we take our short cuts, be they minor infelicities at the margins or major failures of outright fraud. And it is here where we naively convince ourselves and the community at large of the veracity of claims without anything close to principled evidence.

There are a number of current proposals on how to remedy our statistical practices (e.g., Simmons, Nelson, Simonsohn, 2011). Although many of these proposals make sense, most miss the critical point that alternatives must be specified. This is a shame. If the

field is ready to do the hard intellectual work of specifying alternatives, then assuredly better science will result.

References

Abelson, R. P. (1997). On the suprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology*, *75*(4), 772-790. Retrieved from http://psycnet.apa.org/doi/10.1037/0022-3514.78.4.772

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531-533. Retrieved from http://dx.doi.org/10.1038/483531a

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. Retrieved from http://dx.doi.org/10.1037/a0021524

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*(3), 317-335. Retrieved from http://www.jstor.org/pss/2245772

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558-1561.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Dawid, A. P. (2005). Statistics on trial. *Significance*, *2*, 6–8.

de Finetti, B. (1974). *Theory of probability* (Vol. 1). New York: John Wiley and Sons.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204-223. Retrieved from http://www.jstor.org/stable/2958475

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave MacMillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274-290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Quantitative Psychology and Assessment*. Retrieved from 10.3389/fpsyg.2014.00781

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193-242.

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, *100*(469), 1–5.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, *39*, 175-191.

Ferguson, C. J., Rueda, S. M., Cruz, A. M., Ferguson, D. E., Fritz, S., & Smith, S. M. (2008). Violent video games and aggression: Causal relationship or byproduct of family violence and intrinsic violence motivation? *Criminal Justice and Behavior*, *35*(3), 311-332. Retrieved from 10.1177/0093854807311719

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd. Retrieved from http://psychclassics.yorku.ca/Fisher/Methods/

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 69-78.

Forscher, B. K. (1963). Chaos in the brickyard. *Science*, *142*, 339.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate Psi. *Journal of Personality and Social Psychology*, *103*, 933–948.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439-453. Retrieved from `http://psycnet.apa.org/doi/10.1037/a0015251`

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *American Statistician*, *59*, 252-257.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications.* Minneapolis, Minnesota: University of Minneapolis Press.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54–61.

Hauser, M., Cushman, F., Young, L., Jin, R., & Mikhail, J. (2007). A dissociation between moral judgment and justification. *Mind and Language*, *22*, 1-21.

Hill, R. (2005). Reflections on the cot death cases. *Significance*, *2*, 13–15.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157-1164.

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, *15*, 172–181.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* Chichester, United Kingdom: John Wiley & Sons.

Jeffreys, H. (1961). *Theory of probability (3rd edition).* New York: Oxford University Press.

Johnson, V. E., & Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, *107*, 649-660. (PMCID:PMC3867525)

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.

Kruschke, J. K. (2012). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current directions in psychological science*, *5*, 161-171.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology,*, *46*, 806–834. Retrieved from `http://www.psych.umn.edu/faculty/meehlp/113TheoreticalRisks.pdf`

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*, 68-75. Retrieved from `http://dx.doi.org/10.1111/j.2044-8317.2012.02067.x`

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406-419. Retrieved from http://dx.doi.org/10.1037/a0024377

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.12-2.* Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*, 1289-1290.

Myung, I.-J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79-95.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*, 333-380.

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *Journal of the Royal Statistical Society. Series B (Methodological)*, *18*(2), 288-294.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289-337. Retrieved from http://dx.doi.org/10.1007/978-1-4612-0919-5_6

Nobles, R., & Schiff, D. (2005). Misleading statistics within criminal trials: The Sally Clark case. *Significance*, *2*, 17–19.

Nosek, B. A., & Bar–Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Osherovich, L. (2011). Hedging against academic risk. *Science–Business eXchange*, *4*.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Perez, M.-E., & Pericchi, L. R. (2014). Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics and Probability Letters*, *85*, 20-24.

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353.

Pollard, P., & Richardson, J. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*, 159-163.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, *27*, 411–427.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science*, *34*, 311–321.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877-903. Retrieved from `http://dx.doi.org/10.1080/00273171.2012.734737`

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes-factor meta-analysis of recent ESP experiments: A rejoinder to Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*, *139*, 241-247.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374. Retrieved from `http://dx.doi.org/10.1016/j.jmp.2012.08.001`

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225-237. Retrieved from `http://dx.doi.org/10.3758/PBR.16.2.225`

Rozenboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiental information in judgment. *Personality and Social Psychology Review*, *2*, 87–99.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, *55*, 62-71.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992-2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, *136*, 471–485. Retrieved from `http://dx.doi.org/10.1037/a0019457`

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131. (10.1126/science.185.4157.1124)

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149-166.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for ANOVA designs. *American Statistician*, *66*, 104–111.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064.

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*, 298-300.

Author Note

Jeff Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211, rouderj@missouri.edu. This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

Footnotes

[1]The Wikipedia entry, en.wikipedia.org/wiki/Sally_Clark provides an overview her case.

[2]According to Wikipedia, the earliest known written usage of this phrase is from a 1938 *El Paso Herald-Post* newspaper article entitled "Economics in Eight Words."

[3]The extension to the case of unknown variance is well known, see Gönen, Johnson, Lu, & Westfall (2005) and Rouder et al. (2009)

[4]The data in this case are 300 observations with a sample mean of $\bar{Z} = .1635$. The prior is centered at 0 and has a standard deviation of 10.0.

Figure Captions

Figure 1. Critical effect sizes needed to reject the null as a function of sample size for a few rules. The wide, light-color line shows the case for the $p < .05$ rule; the dashed line shows the case for the $\alpha_N = \min(.05, \beta_N)$ rule; the thin solid line shows the case for the $\alpha_N = \beta_N/5$ rule. For the two later rules, an alternative must be assumed, and effect size was set to .4.

Figure 2. Updating with Bayes' rule. **A.** Dashed and solid lines are prior and posterior distributions, respectively; the vertical tick marks near the x-axis are the data. **B.** The change from prior to posterior is how probable the data are for a given parameter value (conditional) relative to how probable the data are on average across ass . **C.** Dashed and solid curves are predicted densities about where the sample mean will occur under the null and general models, respectively. The thin vertical line is the observed sample mean value of the data in Panel A. **D.** The Bayes factor, the predictive density of a sample under the null model relative to under the general model, is shown as a function of sample means. The thin vertical line is for the observed sample mean value, and the ratio, .305 indicates that the null model is about 3/10 as accurate as the general model in predicting the observed data in Panel A..

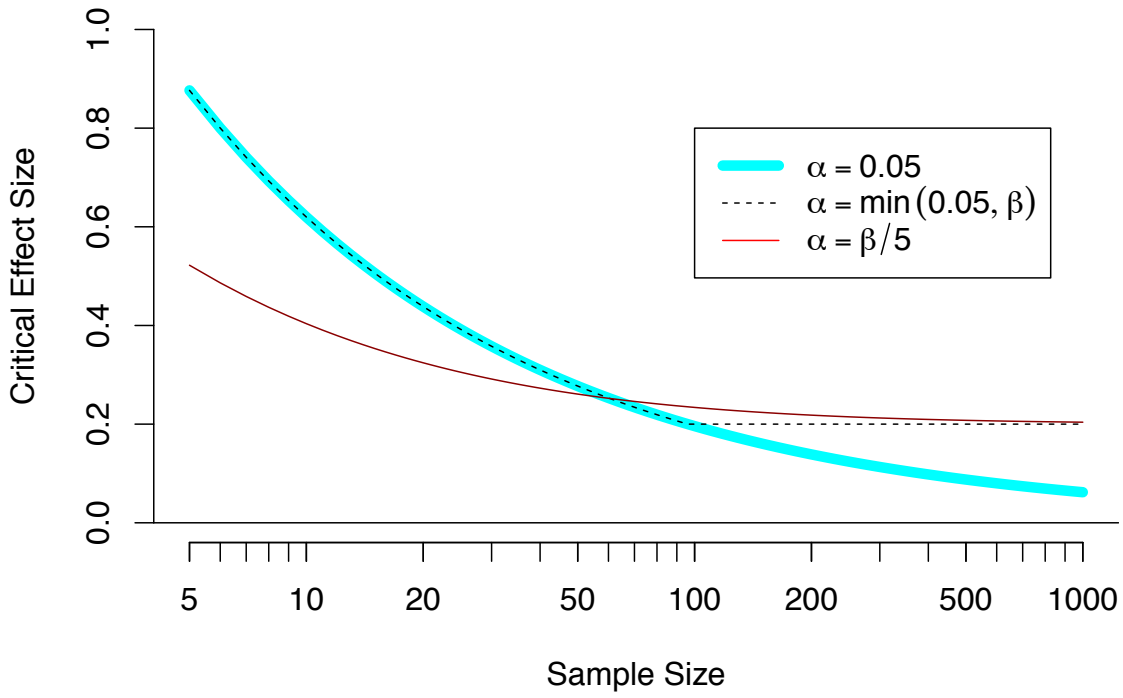
Figure 3. The specification of the alternative affects the Bayes factor. **A.** The null (red arrow) and three alternatives (blue curves). **B.** Predictions for the null model (red) and for the three alternatives (blue). **C.** Bayes factors between the null and each alternative as a function of the observed sample mean.

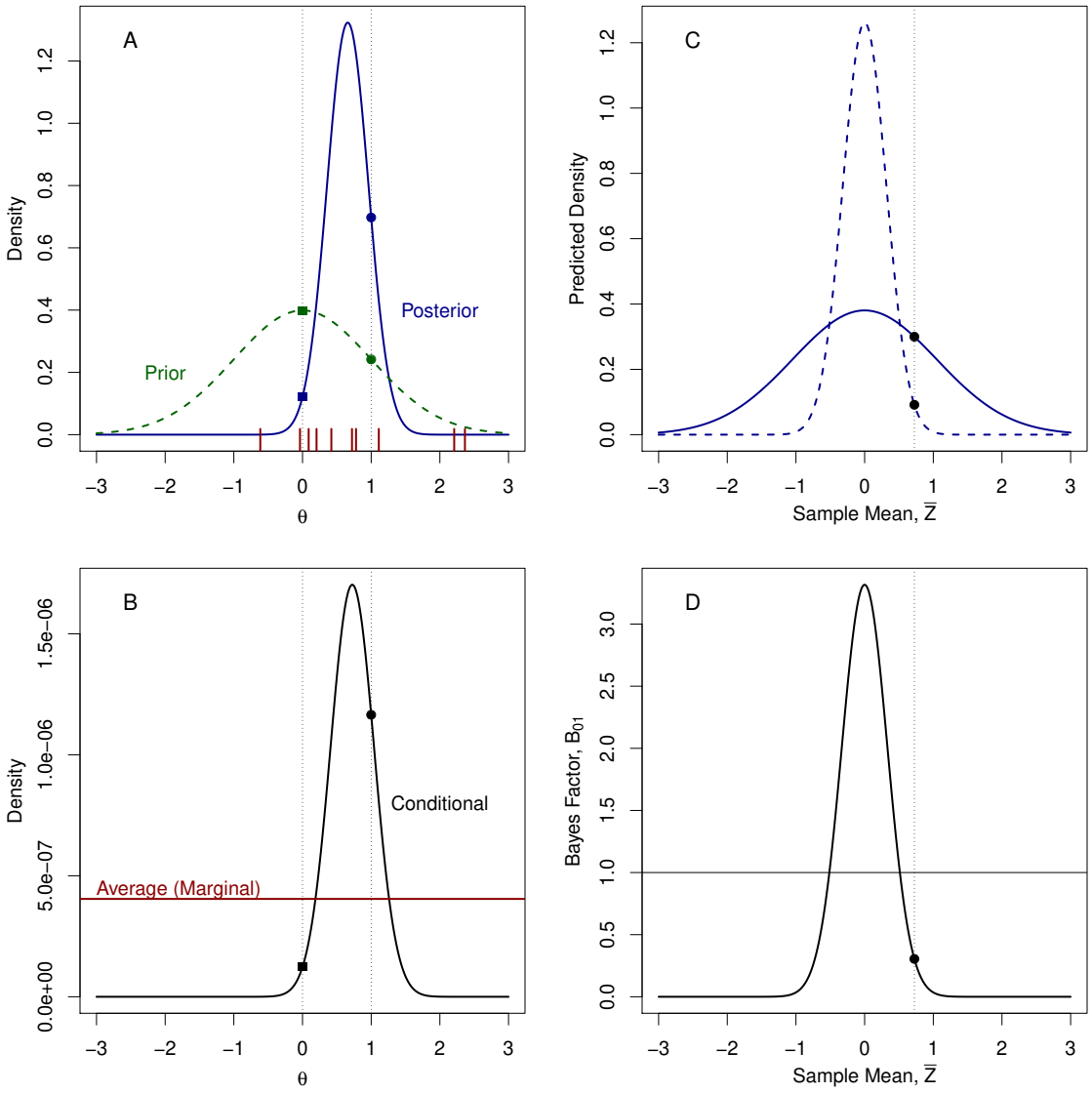
Figure 4. Inference with credible intervals does not satisfy Bayes' rule. The dashed and solid lines are prior and posterior distributions, and the shaded area is the 95% credible interval. This interval is narrow and does not include zero indicating that the null is not

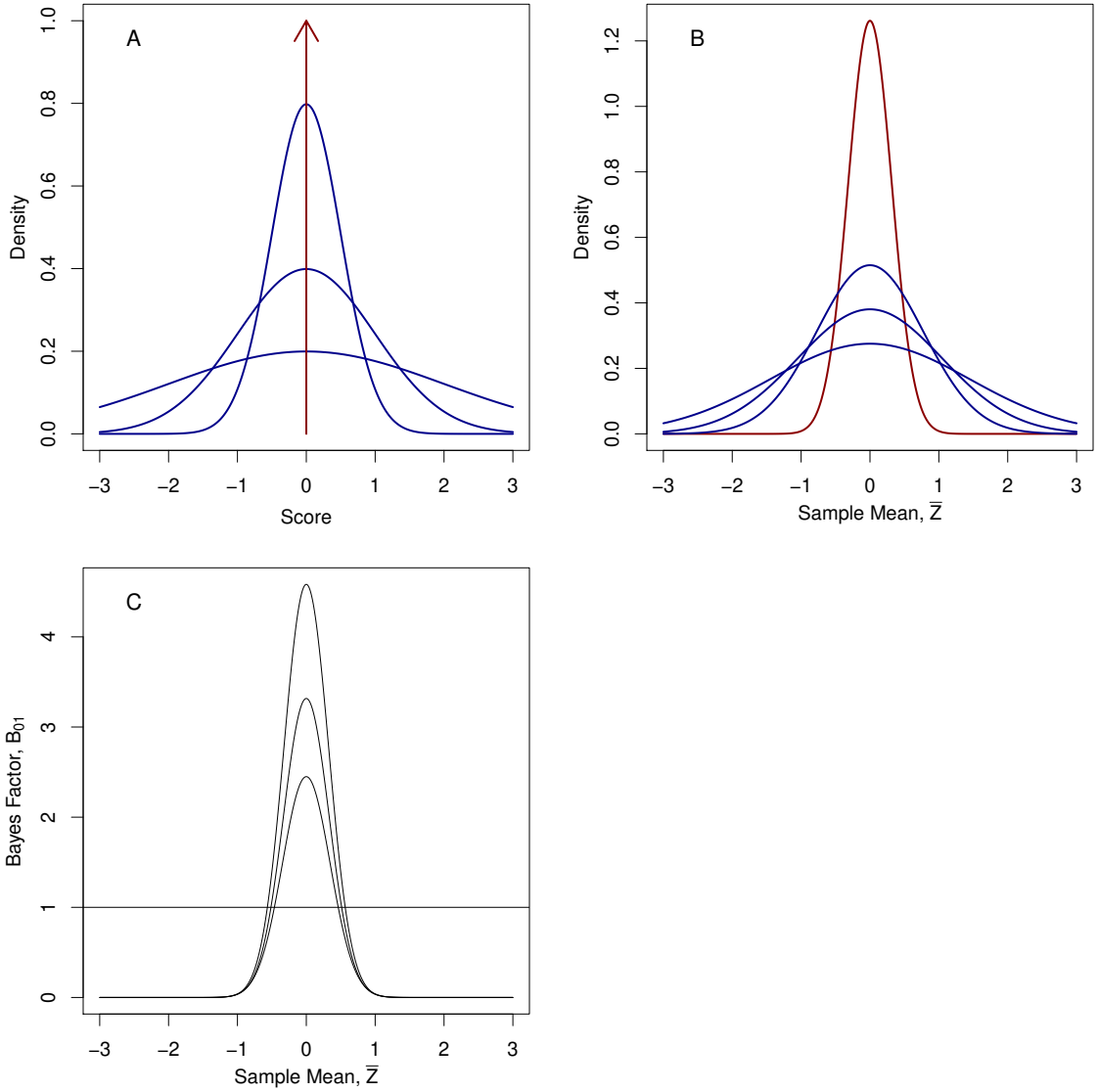
too plausible. Yet, the Bayes factor is 1.0 indicating completely equivocal evidence.

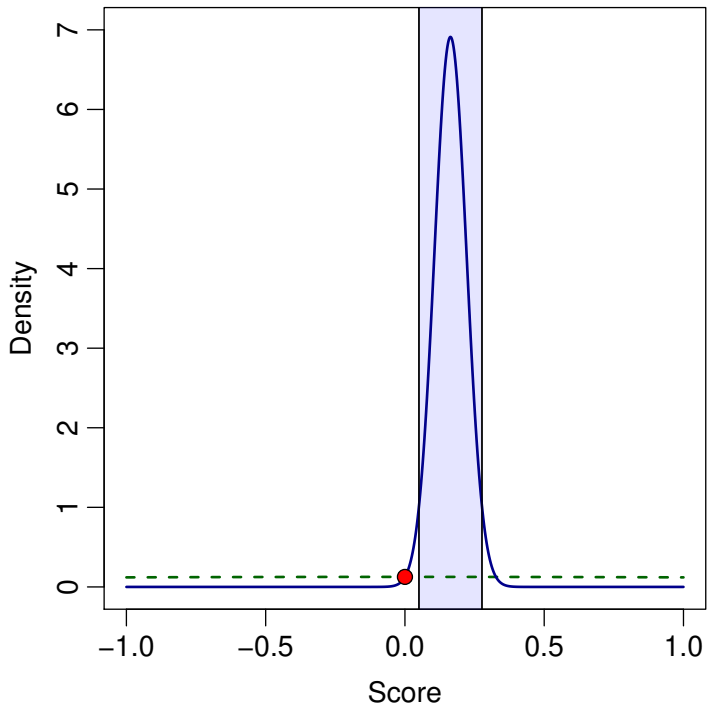
Figure 5. Alternatives may be specified on effect-size measures. **Left.** Four possible specifications. These specifications share the common property of being symmetric around zero and specifying that smaller effect sizes are more common than larger ones. The difference between them is the scale, denoted σ_δ. Scales of 10 and .02, top left and bottom right panels, respectively, are too extreme to be useful. Scales of 1 and .2, top right and bottom left panels respectively, reflect limits of reasonable specifications in most contexts. **Right.** Bayes factor as a function of prior specification (σ_δ) shows that while the specification matters as it must, the changes across reasonable specifications are not great, especially when compared to the sensitivity of the Bayes factor on the data (t-values).

Figure 6. Estimation requires a model. Left: The probability of null model vs. a default alternative as a function of observed effect size for $N = 50$ observations. Right: The best effect size estimate as a function of sample effect size. For large sample effects, the effects model is weighted and the estimate is very close to the sample value. For small sample effects, the null model is weighted and the estimate is shrunk toward zero.

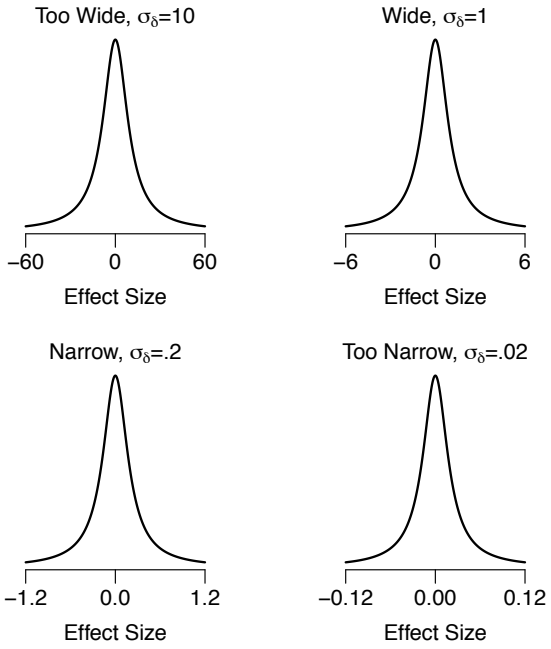

Critical Effect Size
0.0
0.2
0.4
0.6
0.8
1.0
5
10
20
50
100
200
500
1000
Sample Size
α = 0.05
α = min(0.05, β)
α = β/5

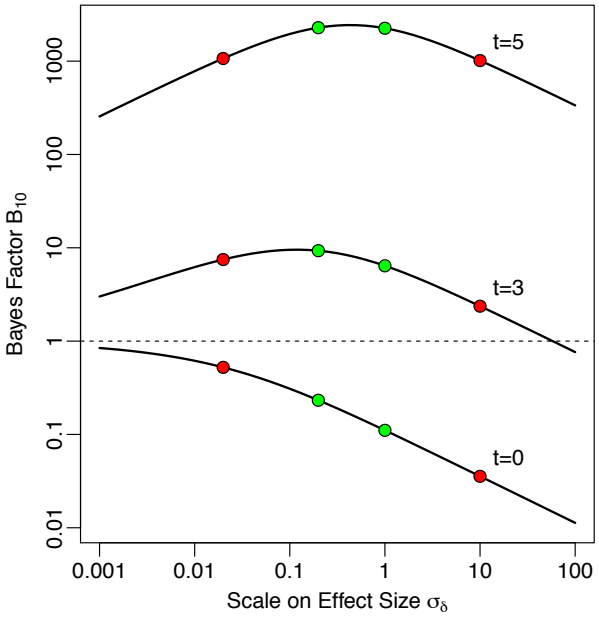

A
Density
θ
Prior
Posterior
B
Density
θ
Conditional
Average (Marginal)
C
Predicted Density
Sample Mean, $\bar{Z}$
D
Bayes Factor, B_{01}
Sample Mean, $\bar{Z}$


A
Density
Score
B
Density
Sample Mean, $\bar{Z}$
C
Bayes Factor, B_{01}
Sample Mean, $\bar{Z}$


Density
0
1
2
3
4
5
6
7
−1.0
−0.5
0.0
0.5
1.0
Score

No Free Lunch, Figure 5





No Free Lunch, Figure 6